Challenge 7

CY6740 – Machine Learning in CyberSecurity

Tejas Krishna Reddy

NUID: 001423166

24th Nov 2020

Apriori and FPGrowth Data Mining Algorithms

Process:

- Read File by lines >> Remove Punctuations >> lower case >> unique words dict >> convert input raw file into corresponding unique numbers dataset.

- Get new_dataset.txt
- Download SPMF.jar file here http://www.philippe-fournier-viger.com/spmf/index.php?link=download.php
- Go to the location in your laptop where spmf.jar file exists
  C:\Users\Tejas\Downloads\MS\Sem4\ML_in_CyberSecurity>
- Run the below line in the command prompt to get the output sheet with 30%+ support!
  java -jar spmf.jar run Apriori new_dataset.txt output.txt 30%
- Run the below line in cmd to get the output sheet for FpGrowth results with minlisft = 1 and minsupport = 50%:
  java -jar spmf.jar run FPGrowth_association_rules_with_lift contextIGB.txt output.txt 50% 90% 1

Results after applying apriori algorithm on new_dataset.txt:

5 item sets that have support value of 30% or higher:

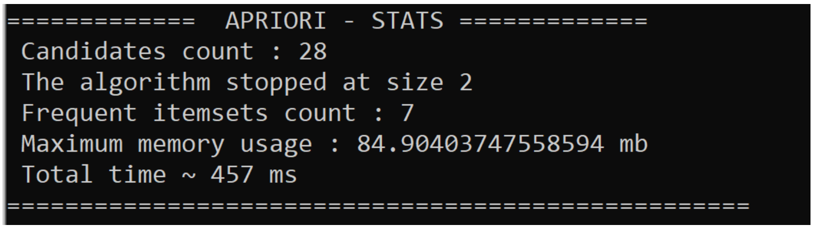
15212 #SUP: 87146

24994 #SUP: 40730

61681 #SUP: 96978

87535 #SUP: 60314

87775 #SUP: 44046



```
=============  APRIORI - STATS =============
 Candidates count : 28
 The algorithm stopped at size 2
 Frequent itemsets count : 7
 Maximum memory usage : 84.90403747558594 mb
 Total time ~ 457 ms
===================================================
```

FPGrowth Algorithm:

A sample of Assosiation rules from the output of fpgrowth algorithm with a minlift = 1 and support > 50%.

61681 ==> 15212 #SUP: 87146 #CONF: 0.9126477949877994 #LIFT: 1.0

15212 ==> 61681 #SUP: 87146 #CONF: 1.0 #LIFT: 1.0

87535 ==> 15212 #SUP: 60314 #CONF: 1.0 #LIFT: 1.0957129415004705

87535 ==> 61681 #SUP: 60314 #CONF: 1.0 #LIFT: 1.0

15212 15212 ==> 87535 #SUP: 60314 #CONF: 0.958643270392269 #LIFT: 1.5176902536715622

61681 87535 ==> 15212 #SUP: 60314 #CONF: 1.0 #LIFT: 1.0957129415004705

15212 87535 ==> 61681 #SUP: 60314 #CONF: 1.0 #LIFT: 1.0

87535 ==> 15212 61681 #SUP: 60314 #CONF: 1.0 #LIFT: 1.0957129415004705

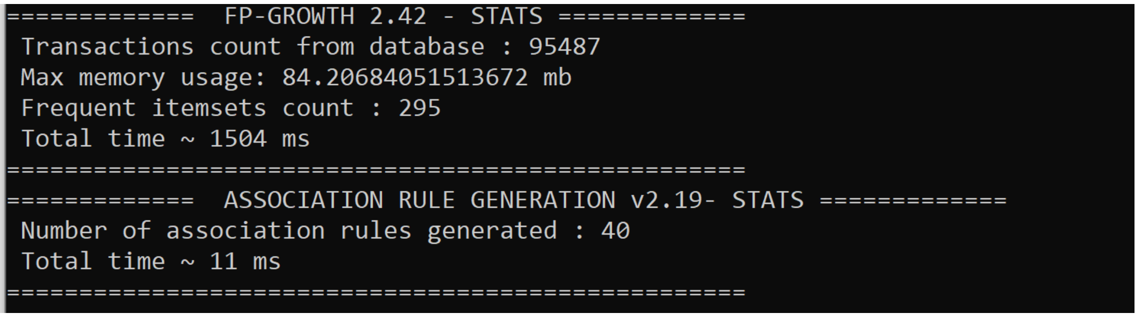
15212 15212 87535 ==> 61681 #SUP: 65160 #CONF: 1.0803461882813277 #LIFT: 1.0803461882813277

15212 15212 61681 ==> 87535 #SUP: 65160 #CONF: 972.5373134328358 #LIFT: 1539.686813140584

15212 15212 ==> 61681 87535 #SUP: 65160 #CONF: 1.035666603089834 #LIFT: 1.6396308805457933

15212 15212 61681 ==> 87535 #SUP: 5399 #CONF: 80.58208955223881 #LIFT: 127.57472535521813

15212 15212 61681 ==> 87535 #SUP: 561 #CONF: 8.373134328358208 #LIFT: 13.256051291772064



```
=============  FP-GROWTH 2.42 - STATS =============
 Transactions count from database : 95487
 Max memory usage: 84.20684051513672 mb
 Frequent itemsets count : 295
 Total time ~ 1504 ms
===================================================
=============  ASSOCIATION RULE GENERATION v2.19- STATS =============
 Number of association rules generated : 40
 Total time ~ 11 ms
===================================================
```

Click here to access new_dataset.txt –
https://drive.google.com/file/d/1XDCefzbvBB1MNPFCxIroN5O4DCIaoaav/view?usp=sharing

Click here to see the Apriori full file results –
https://drive.google.com/file/d/10W1d5pWTpztWlL1mZ2VJsJrUOFnfoM_8/view?usp=sharing

Click here to see the full results of FPGrowth Algortihm -
https://drive.google.com/file/d/1XhBVIXmz1FFHGHnABcYEDjfOUDdDK5N-/view?usp=sharing

Challenge 7 - Apriori and FpGrowth Data Mining Techniques

Author: Tejas Krishna Reddy
NUID: 001423166
Date: 24th Nov 2020

Process:

- Get new_dataset.txt
- Download SPMF.jar file here http://www.philippe-fournier-viger.com/spmf/index.php?link=download.php (http://www.philippe-fournier-viger.com/spmf/index.php?link=download.php)
- Go to the location in your laptop where spmf.jar file exists C:\Users\Tejas\Downloads\MS\Sem4\ML_in_CyberSecurity>
- Run the below line in the command prompt to get the output sheet with 30%+ support! java -jar spmf.jar run Apriori new_dataset.txt output.txt 30%
- Run the below line in cmd to get the output sheet for FpGrowth results with minlisft = 1 and minsupport = 50%:
  java -jar spmf.jar run FPGrowth_association_rules_with_lift contextIGB.txt output.txt 50% 90% 1

In [1]:

```python
# Import necesery modules
import string
```

Read the file:

In [2]:

```python
#Read the file
file1 = open('Dataset_Challenge7.txt', 'r', encoding="utf8")
lines = file1.readlines()

# Split all the strings based on space
all_words = [x.split() for x in lines]
len(all_words)
```

Out[2]:

```
95488
```

Remove Punctuations and convert all words to lower case:

In [3]:

```python
for i in range(len(all_words)):
    all_words[i] = [x.translate(str.maketrans('', '', string.punctuation))

    ## Lower Case
    all_words[i] = [x.lower() for x in all_words[i]]

# Print a sample
all_words[1]
```

Out[3]:

```
['bare', 'tse', 'di', 'woke', 'di', 'tsena', 'mo', 'lockdown']
```

Create a flat list of all words from cleaned lists and create a Dict of unique words:

In [6]:

```python
## create a flat list of all words from cleaned lists
words =  [item for sublist in all_words for item in sublist]

## List of unique words
uwords = list(set(words))

uwords_dict = {k: v for v, k in enumerate(uwords)}
# Print a sample of first 15 items in the dict
list(uwords_dict.items())[1:15]
```

Out[6]:

```
[('hula', 1),
 ('stando', 2),
 ('hahaahahahaha', 3),
 ('88253cr', 4),
 ('dell,äôisola', 5),
 ('rightu', 6),
 ('antivaxxers', 7),
 ('httpswwwfacebookcom100013345004443posts994317527689738sfnsnwiwspmoampext
idqi32xw8ppj0nmdvu',
  8),
 ('thanelockdown', 9),
 ('pre57', 10),
 ('investigated', 11),
 ('tjeeer', 12),
 ('resenting', 13),
 ('mindfuck', 14)]
```

Convert the input text into corresponding numbers and save new_dataset.txt

In [ ]:

```
myfile = open('new_dataset.txt', 'w')
for i in range(len(all_words)):
    current_line = all_words[i]
    wordNum = []
    for x in current_line:
        myfile.write("{} ".format(uwords_dict[x]))

    myfile.write("\n")

myfile.close()
```

In [ ]:

In [ ]: